Properties of evolving e-mail networks

Juan Wang and Philippe De Wilde
Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, United Kingdom
(Received 13 January 2004; revised manuscript received 17 June 2004; published 10 December 2004)

Computer viruses spread by attaching to an e-mail message and sending themselves to users whose addresses are in the e-mail address book of the recipients. Here we investigate a simple model of an evolving e-mail network, with nodes as e-mail address books of users and links as the records of e-mail addresses in the address books. Within specific periods, some new links are generated and some old links are deleted. We study the statistical properties of this e-mail network and observe the effect of the evolution on the structure of the network. We also find that the balance between the generation procedure and deletion procedure is dependent on different parameters of the model.

DOI: 10.1103/PhysRevE.70.066121 PACS number(s): 89.75.Da, 89.20.Hh, 89.75.Fb

I. INTRODUCTION

In the past decades, the structures of various networks in the real world have been well studied by many researchers. Erdös and Rènyi first introduced random-graph theory in 1959 [1], in which edges are distributed randomly and the presence or absence of any edge between two nodes in the network is dependent on a fixed connection probability p. Some further mathematical development of random-graph theory is described in [2]. It has been found that the degree distribution of a random graph networks follows a Poisson distribution,

$$P(k) = \binom{n}{k} p^k (1-p)^{N-k} \simeq \frac{z^k e^{-z}}{k!}, \quad k \geqslant 0.$$

Numerous networks, particularly in epidemiology studies, have been viewed and analyzed as random graphs. However, random graphs fail in describing the structural properties of some real-world networks. The study of networks such as networks of movie actor collaboration [3,4], science collaboration [5], WWW [6,7], and Internet [8] found that the degree distribution of these networks deviates measurably from a Poisson distribution, but follows a power-law distribution: $P(k) \sim k^{-r}$. It indicates that the networks can self-organize to a scale-free state, where some highly connected "hub" nodes strongly affect the structure and dynamics of the networks. In 1999, Barabàsi and Albert presented in [4] that the origin of this scale-free behavior was found to be a consequence of two mechanisms: growth of nodes and preferential attachment to well connected nodes.

The understanding of topological and statistical properties of these networks is becoming very important. The methods for measuring a network's topology such as degree distribution, average path length, and clustering coefficient have been illustrated in [9–11].

Networks with scale-free behaviors are robust to the failure of some node. In contrast, they are very vulnerable to attack since the direct attack to "hub" nodes could cause severe damage to the entire network. Therefore, currently, the spread of infection becomes of special interest in the study of complex networks. Some research has been done to investigate how the connectivity of these networks affects the spread of human diseases and computer viruses [12,13].

Currently, many computer viruses spread through a network by attaching to an e-mail message and sending themselves to many other people whose e-mail addresses are in the recipient's e-mail address book, a file containing a list of e-mail addresses of frequent correspondents. This structure does not have to be the address book as implemented in most e-mail software, it can also just be a file with e-mails that are used to send new e-mails.

Some research work [14,15] used real data from server log files or address books of large computer systems and applied the data in computing the statistical properties. Their results show that the e-mail networks display scale-free and small-world behaviors, indicating that viruses are easy to spread in real e-mail networks.

In our studies, our emphasis is different from other research work: (i) we construct an e-mail network model that is intermediate in sophistication between the simplified models needed for application of statistical physics methods, and models of the real world that require numerous parameters; (ii) in the line of our previous work on dynamic behavior of a network in [16], we regard our e-mail address network model as an evolving network, in which links can be added and deleted periodically based on some rules. This is different from the two mechanisms in Barabàsi's scale-free model [4].

In [14], it is shown that removing some suitably selected vertices or disabling the account of the most connected node can slow down the spread of a virus through the network and save some time for a patch. In this paper, we assume users in the network organize their e-mail address book periodically by deleting and adding some e-mail addresses of users. In this paper, we perform an analysis of our model. We also present the results of our simulation with different parameter settings to analyze the effects of evolution on the structure of the e-mail address network.

II. THE EVOLVING E-MAIL ADDRESS NETWORK MODEL

E-mail networks are quite different from the networks mentioned above in that some social networks were modeled as bipartite graphs and some as undirected graphs, while e-mail networks are directed graphs, indicating that each link

in the network has a direction. In our model, the nodes of the e-mail network represent e-mail address books of different users, which are connected by links running from user A to user B if B's e-mail address is in A's e-mail address book.

We construct an e-mail network with N nodes, and in order to investigate the scalability of the network, the size of the network is increased in different simulations. We attempt to create a simple virtual e-mail exchange network as follows: at each time step, we assume that an individual user in the network sends a specific amount of e-mails to all the users in the network. We denote the number of e-mails sent from the source node i to the target node j by k_{ij}. Then the total number of e-mails E exchanged in the network within each time step is

$$E = \sum_{i=1}^{N} \sum_{j=1}^{N} k_{ij}, \quad k_{ij} = 0,1,2, \ldots . \tag{1}$$

Here, users are allowed to send e-mails to themselves because people usually use this to save or transfer some files, or to test whether their e-mail boxes are working. Users who receive these e-mails from the sender are selected randomly in the network.

At the initialization stage ($t=1$), the network is built up by connecting every source node to all its target nodes with an individual link if at least one e-mail has been sent from the source node to the target nodes. The connectivity of this network can be represented by an $N \times N$ *adjacency matrix* C. The value of an element C_{ij} is either one or zero, indicating that there is a link running from source node i to target node j or vice versa,

$$C_{ij}(1) = \begin{cases} 1, & k_{ij}(1) \geqslant 1, \\ 0, & k_{ij}(1) = 0. \end{cases} \tag{2}$$

After the initialization stage, the e-mail network evolves following the rules of generation and deletion.

Generation. We let all users in the network check the amount of e-mails that have been sent from them to other users every t_g time units. Here t_g indicates how often we carry out the generation procedure. Within period t_g, if the number of e-mail contacts from source node i to target node j exceeds the generation threshold g, node j's e-mail address should appear in node i's e-mail address book, corresponding to a link running from i to j. In this case, we add a link from i to j if there is no link existing. On the other hand, if there is already a link existing, the connectivity will not change. Thus the connectivity after the generation procedure can be presented as

$$C_{ij}(t + t_g) = \begin{cases} C_{ij}(t) + \Delta C, & \Delta k_{ij}(t_g) > g, \\ C_{ij}(t), & \Delta k_{ij}(t_g) \leqslant g, \end{cases} \tag{3}$$

where $\Delta C = (1 - C_{ij})$, indicating that when $C_{ij}(t)$ is zero (no link existing), add 1 to the new value of C_{ij} and when $C_{ij}(t)$ is 1 (link exising), add 0 (no change) to the new value. The $\Delta k_{ij}(t_g)$ is the number of e-mails sent from source note i to target node j within period t_g, $\Delta k_{ij}(t_g) = k_{ij}(t+t_g) - k_{ij}(t)$.

Deletion. As discussed in Sec. I, we can clean up the e-mail address book periodically to slow down the spread of viruses through the network. Within every restricted period t_d, users check the amount of e-mails sent from them to other users; if they find that the amount to some users whose addresses are already in their address books is less than the deletion threshold d, they delete these e-mail addresses. This corresponds to the situation when the links running from source nodes to the seldom contacted nodes are no longer existing in the network. Thus the connectivity of nodes in the network after the deletion procedure can be shown as

$$C_{ij}(t + t_d) = \begin{cases} C_{ij}(t), & \Delta k_{ij}(t_d) > d, \\ 0, & \Delta k_{ij}(t_d) \leqslant d, \end{cases} \tag{4}$$

where $\Delta k_{ij}(t_d)$ is the number of e-mails sent from source node i to target note j within period t_d, $\Delta k_{ij}(t_d) = k_{ij}(t+t_d) - k_{ij}(t)$.

By investigating the connectivity C_{ij}, we can further study how these parameters illustrated above such as g, d, t_d, t_g affect the structural properties of the evolving e-mail network, e.g., average number of links in the network, average path length, clustering coefficient. This will be shown in Sec. IV.

Recently, some studies [17,18] show that it is important and feasible to investigate statistical properties of complex networks by assigning weights to edges as in complex networks. Particularly, [17] shows how the strength of the relations in e-mail networks can be measured. Moreover, the mechanism called preferential exchange based on the idea of positive feedback has been found suitable to model e-mail networks. For our model, it is a possible direction for further studies to assign weights of edges according to the amount of e-mail exchanged over links, so that some other network properties, such as vertex strength in [18] which is in terms of weights and adjacency matrix, can be investigated just as the structural properties that we have studied in this paper. We can also study the effects of parameters correlating with generation and deletion procedures discussed in the paper on structural properties which incorporate the weights of the connections. Moreover, comparison of results can be made between models using preferential exchange and preferential attachment.

III. ANALYSIS OF THE MODEL

Erdös discovered that the probabilistic method can be used in solving problems in graph theory. Some research presents methods to calculate the edge probabilities as in [19].

According to the description in Sec. II, in our model the probability of obtaining a direct link from one node to another node depends on whether there are enough e-mails sent from source node to target node. In addition, the threshold of generating links g and deleting old links d, the generation interval (t_g), and deletion interval (t_d), also play very important roles on the evolution of e-mail networks.

We model the e-mail sent from source node i to target node j as the event where node i selects some target nodes among all the nodes in the network. In this way, the analytical study of our model can be started by computing the

probability of each node being selected by node i for k_{ij} times, $0 \leqslant k_{ij} \leqslant k$. We use a statistical method named generalized Bernoulli trials [20], which is used to calculate the probability of events occurring different times to address the problem.

The size of the network is N, and at each time step, every source node sends k e-mails to other nodes. Some nodes can be selected as target node more than once, representing that a user can send more than one e-mail to any user, even itself. We denote the event in which each individual node is selected as receiver by $a_1, a_2, \ldots, a_N$ with

$$p_1 = p_2 = \cdots = p_N = \frac{1}{N} \tag{5}$$

in our model. Therefore, as in the case of generalized Bernoulli trials, the probability of the event $\{a_1$ occurs k_1 times, a_2 occurs k_2 times ,…, a_N occurs k_N times$\}$ is

$$P_n(k_1, k_2, \ldots, k_r) = \frac{n!}{k_1!\, k_2!\, \cdots k_N!} \left(\frac{1}{N}\right)^n, \tag{6}$$

where n is the amount of e-mails sent from node 1 to others within a specific period.

Thus, if we know the values of $\{k_1, k_2, \ldots, k_N\}$, the probability in Eq. (6) above can be calculated, so that the probability of one node having m outgoing links can be further studied.

We have written a program to get all the combinations of vectors k for every specific m. However, we found that the computation time required to calculate them is exponential in the size of the network, making it impractical for further analysis.

IV. SIMULATION RESULTS

We now investigate the effects of different parameter settings on the topology of our model, by simulation. We construct the e-mail address network according to the model description in Sec. II.

A. Case I: Equivalent e-mail contact

First, we assume that at each time step, user i sends $k_i = k = 20$ e-mails to other randomly selected users. Although this assumption may not be an exact scenario in the real world, since a person with more e-mail addresses will probably send more e-mails than a person with a small e-mail address book, this simplification is made as a starting point of our analysis. In Sec. IV B, we will illustrate the case in which e-mails are sent from user i, k_i relates to the degree of node i, or the size of user i's address book.

1. Average number of links

The simulations for investigating the average number of links ran for 2000 time steps and the number of nodes in the network is set to $N = 1000$. We denote the number of links connected to node i by E_i, thus the average number of links for the whole network E is

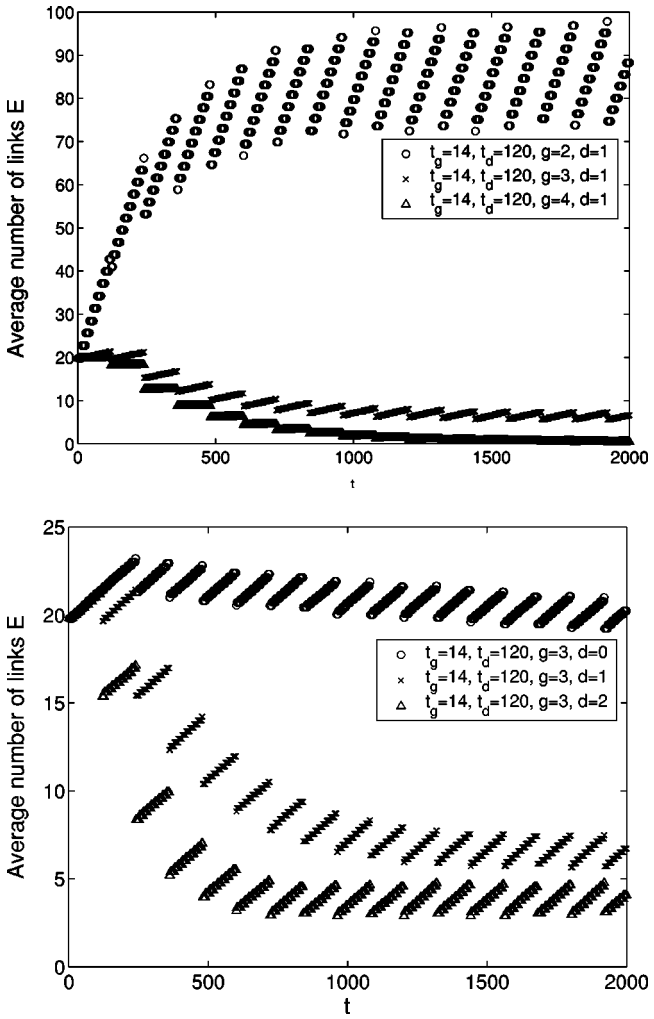

FIG. 1. Effect of generation threshold g and deletion threshold d on the average number of links E, with a network size of $N = 1000$.

$$E = \frac{1}{N} \sum_i E_i. \tag{7}$$

The result of how the generation threshold g and deletion threshold d affect E is shown in Figs. 1(a) and 1(b), respectively. We carried out numerous simulations keeping generation interval $t_g = 14$ and deletion interval $t_d = 120$ the same in both studies of g and d, while in order to investigate the effect of generation threshold g, we fixed $d = 1$ with varying g ($g = 2, 3, 4$), and to study the effect of deletion threshold d, the generation threshold g was set to $g = 3$ with varying d ($d = 0, 1, 2$).

In Fig. 1, we find that the results agree well with the assumptions of our model, as illustrated in Sec. II. The graphs show a sawtooth pattern. This can be explained as follows. The average number of links E keeps increasing with the continuous insertion in the generation procedures for a period t_g. After the deletion procedure is carried out every t_d time steps, the E shows a sudden drop. The total width of one "tooth" of the sawtooth pattern is t_d. Thus the pattern of E is not an implication of discontinuities, but the

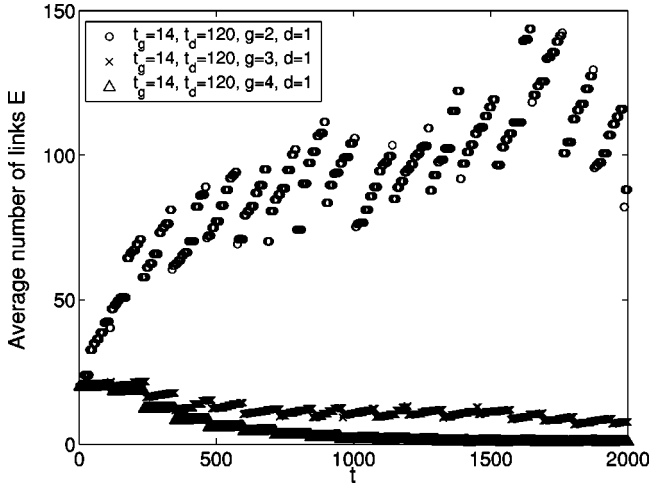

FIG. 2. Effect of generation threshold g on the average number of links with Poisson distribution and a network size of $N=1000$: $g=2(\circ),3(\times),4(\triangle)$.

result of the increase and decrease of E caused by the generation and deletion procedures. To further verify this, instead of implementing discrete t_g and t_d, we have also run simulations with generation and deletion interval following a Poisson distribution for each user, such that the average is t_g and t_d. The results are presented in Fig. 2. We can see that the form of E in Fig. 2 is similar to that in Fig. 1. But it shows some slight difference in which E varies irregularly within a range in Fig. 2 instead of a clear "sawtooth" pattern in Fig. 1.

Furthermore, we can observe that in Fig. 1(a), the average number of links in the network E with smaller g is always larger than with bigger g, where links are generated by the continuous generation procedure according to g and deleted by the deletion procedure according to d. This can be explained as follows: in this set of simulations, although source nodes send the same amount of e-mails to some specific nodes, it is more difficult for source nodes to generate links to target nodes with the same deletion threshold and higher generation threshold.

We also find in Fig. 1(a) that, as time elapses, the tendency of E reflects the relationship between the generation and deletion effect on the network. If the generation procedure effect is stronger than the deletion procedure as when $g=2(\circ)$, the tendency of E is increasing. In contrast, when the generation threshold is high, as $g=3(\times)$ and $g=4(\bullet)$, so that the deletion effect is stronger, E tends to decrease. Finally, the two procedures balance each other, and the time average (over a period larger than t_d) of E in the network becomes constant.

In Fig. 1(b), we can see that with three parameters (t_g, t_d, and g) fixed, E with smaller d is larger than with bigger d, indicating that smaller d causes links between source node and target node more likely to be deleted [refer to Eq. (4)]. Comparing (b) with (a), the deletion procedures shown in the three graphs in (b) seem to dominate the process more than the generation procedures. Especially with bigger d, as $d=1(\times)$ and $d=2(\triangle)$, E decreases drastically at the early stage of simulation and stays in the stable region after $t=600$, indicating that the two effects balance each other very quickly.

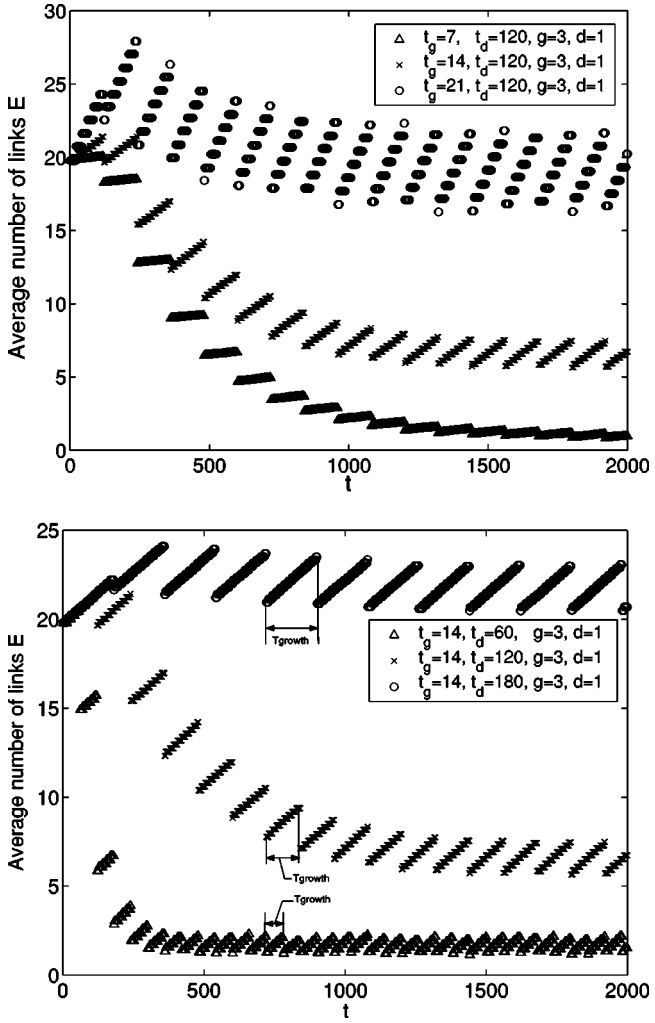

FIG. 3. Effect of generation time interval t_g and deletion time interval t_d on the average number of links E, with a network size of $N=1000$.

In Fig. 3, we show the effect of generation time interval t_g and deletion time interval t_d on E. In order to investigate these effects, we fixed some parameters as follows: (i) for the study of generation interval t_g [Fig. 3(a)], we set $g=3$, $d=1$, and $t_d=120$ with varying t_g ($t_g=7$, 14, and 21); (ii) for the study of deletion interval t_d [Fig. 3(b)], we fixed $g=3$, $d=1$, and $t_g=14$ with varying t_d ($t_d=60$, 120, and 180).

The evolution of E in Fig. 3 is consistent with the settings of time interval t_g and t_d. We can observe that in (a), with the same t_d, E increases as t_g increases, and similarly in (b), with the same t_g, E increases as t_d increases. This is because within longer period (bigger t_g and t_d), generally, the number of e-mails sent from one node to another is more than within a shorter period so that it is easier to fulfill the requirement of generating new links and, on the other hand, harder to fulfill the requirement of deleting links. Furthermore, we observe that more apparently in (b), the time interval T_{growth} of continuous growth of E decreases as t_d decreases. This can be explained as follows: as described in Sec. II and above, the generation and deletion procedures are carried out in an interlaced manner with asynchronous t_g and t_d. If the deletion

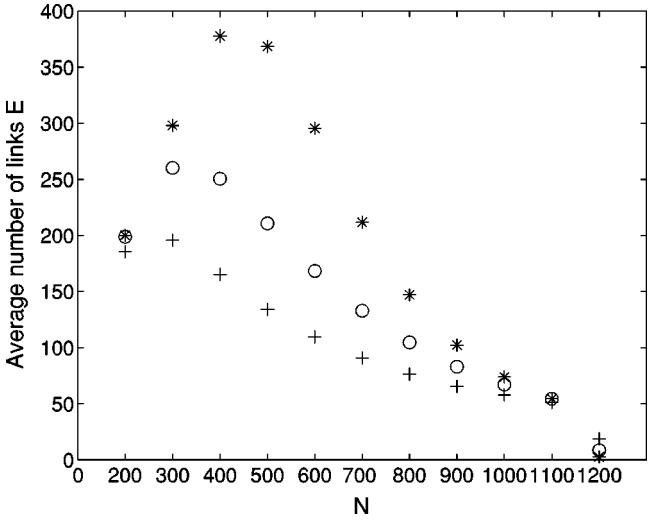

FIG. 4. Average number of links E vs network size N at different time steps $t=200$ (+), $t=300$ (∘), and $t=2000$ (*). Parameter settings: $t_g=14$, $t_d=120$, $g=2$, and $d=1$.

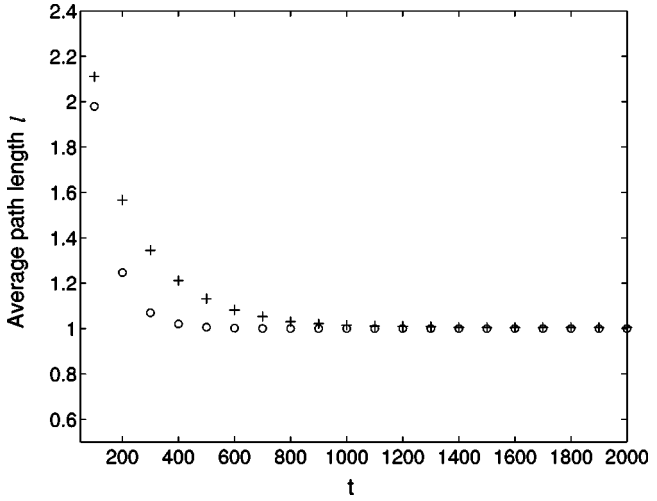

FIG. 5. Average path length ℓ vs time t, with different network sizes of $N=200$ (∘) and $N=300$ (+). Parameter settings: $t_g=14$, $t_d=120$, $g=2$, and $d=1$.

procedure is executed more frequently (smaller t_d), T_{growth} is forced to be shorter.

We now investigate the scalability of our e-mail address network. Figure 4 shows E versus different numbers of nodes in the network N at $t=200$ (+), $t=300$ (∘), and $t=2000$ (*), respectively, with $t_g=14$, $t_d=120$, $g=2$, and $d=1$.

Some evolution characteristics of the network are displayed in Fig. 4: for one specific N, E increases as time elapses (increasing t, along the vertical direction for each N from "+" to "∘" to "*"). We find that the speed of increase of E is faster in time interval [200, 300] than in [300, 2000], indicating that the network evolves more quickly in the earlier stage of evolution. Moreover, for small size networks such as $N=200$, 300, and 400, we can see that E reaches its saturated status at the end of evolution, which means that the network can easily become fully connected when there are only a few nodes in the network. However, for a bigger size of the network as $N>400$, the network cannot reach fully connected status anymore, and as N increases, the increase of E decreases and even changes to a minor increase after $N=1100$. This result is consistent with the assumption of equivalent e-mail contact between users, resulting in users in big size networks usually receiving fewer e-mails than in small size networks.

2. Average path length

There is no closed formula to compute the average path length ℓ yet. But it is widely accepted that this ℓ follows some scaling form as a function of a network model's parameters, e.g. size of network N, connection probability p, and so on.

In our simulation, we use the Dijkstra algorithm to compute the shortest path length ℓ_{ij} between any two nodes: node i and node j. From this we can obtain the average path length for the whole network as

$$\ell = \frac{1}{N^2}\sum_i \sum_j \ell_{ij}, \tag{8}$$

where $i,j=1,2,\ldots,N$. We fixed parameters as $t_g=14$, $t_d=120$, $g=2$, and $d=1$ with different network sizes of $N=200$ (∘) and $N=300$ (+). The result of ℓ versus t with this setting is shown in Fig. 5.

It is shown that with $N=200$ and $N=300$, ℓ monotonously decreases as time is elapsing until $\ell=1$. The $\ell=1$ implies that every node is directly connected to all other nodes in the network, indicating the network is fully connected. Average path lengths of different networks have been studied, and are summarized in [11]. The average path length of "1" of our model has not been confirmed by previous empirical research. However, it is consistent with the algorithm used in our model. At each time step, every node is assumed to send k e-mails to other nodes in the network; in the cases of a small size network such as $N=200$ and $N=300$ here, most of the nodes are very likely to receive many e-mails, which results in generation of links happening very frequently (according to Sec. II). This makes our e-mail network become fully connected with $\ell=1$ after a period of evolution. Moreover, we also think that because of the evolution of technology, space and memory for personal e-mail address books become bigger so that people always keep almost all of the contact addresses in their address books. This indicates that one node is very likely to connect directly to all other nodes, resulting in a fully completed graph with the average path length equal to "1".

Furthermore, we found that the value of $\ell_{N=300}$ (+) is bigger than the value of $\ell_{N=200}$ (∘) before they reach 1 and it takes longer for $\ell_{N=300}$ than $\ell_{N=200}$ to reach 1. This result is in accordance with the result shown in Fig. 4 under the same setting, in which at the beginning of the evolution of the network $E_{N=200}\simeq 180$ is closer to 200, the value of N, compared with $E_{N=300}\simeq 200$ to 300. The average number of links E increases until $E_{N=200}=200$ and $E_{N=300}=300$ become fully

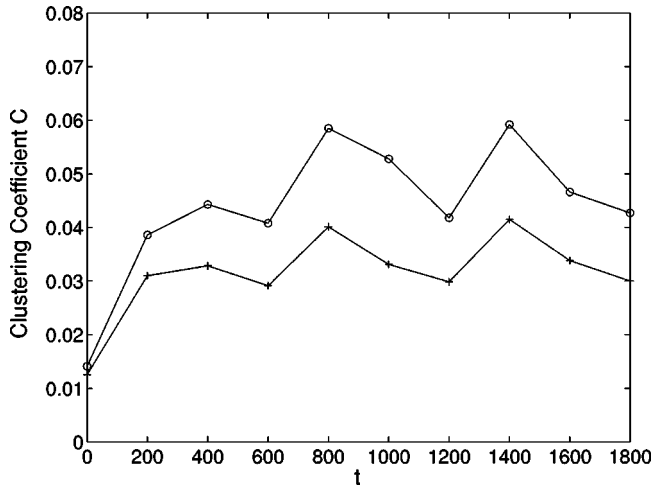

FIG. 6. Clustering coefficient C vs time t, with different network sizes of $N=1000$ (∘) and $N=1100$ (+). Parameter settings: $t_g=14$, $t_d=120$, $g=2$, and $d=1$.

connected. Here we have not studied ℓ for large size networks, because it takes a very long time to compute the shortest path between any two nodes for the whole network. But based on the analysis of results shown in Fig. 4 and Fig. 5, we may predict that, for a large size of networks, the average path length would increase as N increases, thus it cannot reach "1" as small size networks do.

3. Clustering coefficient

Clustering is another important property of networks to be investigated. The definition of it is provided by the fraction of fully connected triples (triangles) to the number of triples of vertices in the network [21]. Thus the clustering coefficient C can be calculated by

$$C = \frac{3 \times (\text{number of triangles in the network})}{(\text{number of connected triples of vertices})}. \quad (9)$$

In our studies, we use an alternative definition of C in [3,11] because it is easier to calculate on a computer for simulation,

$$C_i = \frac{(\text{number of triangles connected to vertex } i)}{(\text{number of triples centered on vertex } i)}. \quad (10)$$

Then the average clustering coefficient C for the whole network is (summation over all vertices i)

$$C = \frac{1}{n}\sum_i C_i. \quad (11)$$

To study C, we set the parameters of our simulation as $g=2$, $d=1$, $t_g=14$, and $t_d=120$. We compare the C of different size networks as $N=1000$ and $N=1100$. The result is shown in Fig. 6.

We observe that the clustering coefficient C fluctuates, but within a range. The value of C for smaller size networks $C_{N=1000}$(∘) is always bigger than for larger size networks $C_{N=1100}$(+). In [9,22], the studies on the statistical properties of a network as a function of the network size N show that

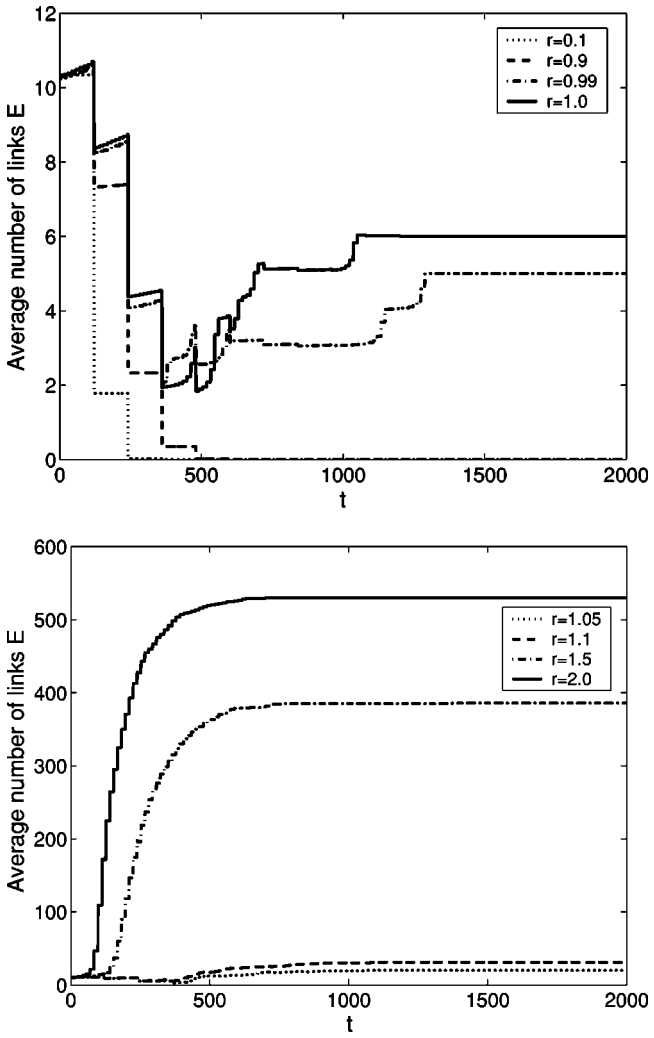

FIG. 7. Effect of contact ratio r on the average number of links E, with a network size of $N=1000$. Parameter settings: $t_g=14$, $t_d=120$, $g=2$, and $d=1$.

for both of the random graph model and the Barabási-Albert model, as the network size N increases, the average path length increases while the clustering coefficient decreases. For our model, results presented in Fig. 5 and Fig. 6 are in accord with the indications. In Fig. 5, the average path length of $N=200$ is smaller than ℓ of $N=300$. On the other hand, the clustering coefficient of $N=1000$ is larger than C of $N=1100$ as in Fig. 6.

Here, we have studied C on a large size network, and we can make a prediction that the clustering coefficient of small size networks would be larger than the results shown in Fig. 6.

The time averages of C are $\overline{C}_{N=1000}=4.41\times10^{-2}$ and $\overline{C}_{N=1100}=3.14\times10^{-2}$, which are similar to the results shown in [15].

This indicates that our e-mail network also has a characteristic property of high clustering, namely as a small-world network.

B. Case II: Degree-related e-mail contact

Having investigated the case of an equivalent number of e-mail contacts between every individual node and other

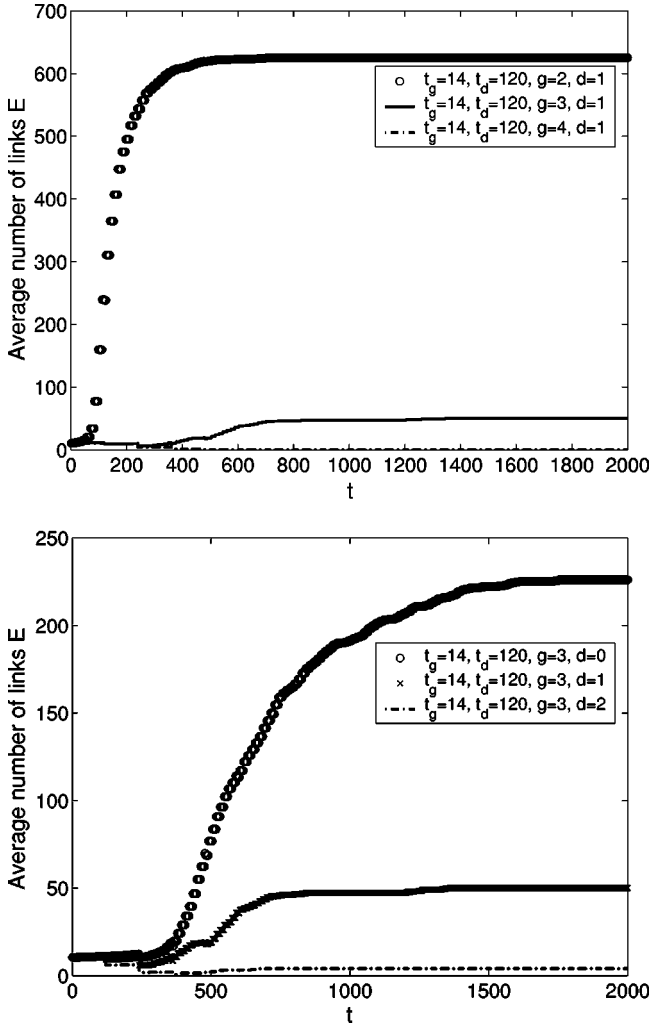

FIG. 8. Effect of generation threshold g and deletion threshold d on the average number of links E, with a network size of $N=1000$.

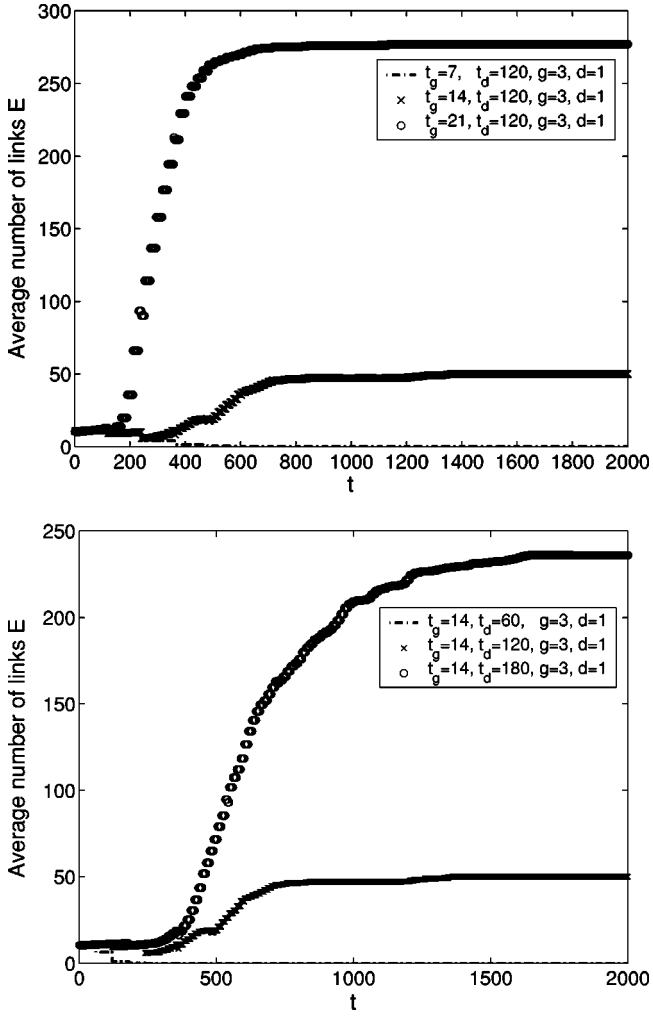

FIG. 9. Effect of generation threshold t_g and deletion threshold t_d on the average number of links E, with a network size of $N=1000$.

nodes as in case I, we now study the case that at each time step, the number of e-mails sent from one node to others relates to the size of the user's address book. The amount of e-mail sent from user i at each time step k_i is controlled by a constant contact ratio r. Therefore, we define k_i as $k_i(t+1)=rE_i(t)$, where E_i is the degree of user i, or the size of user i's address book.

First, we investigate the average number of links E with different contact ratio r. Figure 7 shows the results of simulations with $t_g=14, t_d=120, g=3$, and $d=1$, (a) for $r\leqslant 1$, and (b) for $r>1$, respectively. We find that E is much smaller for $r\leqslant 1$ than for $r>1$. In (a), when $r\leqslant 0.9$, the average number of links vanishes gradually as time elapses. For the case of $r=0.99$ and $r=1$, although E decreases at the beginning, after a time period, the tendency of decrease stops and nodes have several connections in the network. In (b), E increases steadily as time elapses until it reaches some value and remains constant afterwards. This is more noticeable for bigger values of r such as $r=1.5$ and $r=2.0$.

Through large numbers of simulations with different parameters, we find it is observable that there exists an absorbing state transition located at a specific value of r, so that for $r\leqslant R$, links in the network vanish gradually. On other hand, for $r>R$, links increase progressively and the number of links becomes constant at the end, as the results show in Fig. 7. However, the location of the absorbing state transition is not fixed for simulations with different parameters. For the simulation presented in this paper, it is located at $r=R$, where $0.9<R<1$. We conjecture that the value of R correlates with settings of parameters.

In order to study the effects of g, d, t_g, and t_d on E in this case, we set the contact ratio $r=1.1$, so $k_i(t+1)=1.1E_i(t)$ instead of $k_i=20$, while keeping other parameters the same as in case I as follows: (i) for study of g, $t_g=14, t_d=120, d=1$, and $g=2,3,4$; (ii) for study of d, $t_g=14, t_d=120, g=3$, and $d=0,1,2$; (iii) for study of t_g, $t_d=120, g=3, d=1$, and $t_g=7,14,21$; (iv) for study of t_d, $t_g=14, g=3, d=1$, and $t_d=60,120,180$. Simulation results are shown in Fig. 8 and Fig. 9.

In Fig. 8 and Fig. 9, we can see that the overall value of E in this case is much larger than in case I, but it is because of the high contact ration that we have chosen. Moreover, we observe that instead of the sawtooth pattern shown in graphs

in case I, E appears to be more smooth in this case. Through investigation of the amount of links having been generated by the generation procedure at every t_g, and links having been deleted by the deletion procedure at every t_d, we find that the smoothness of E is a result of the fact that there is always only one of the two procedures dominating the network so that the sawtooth pattern disappears, and we also find that both of the two procedures play a lesser role on the network when it is in the later stage of evolution. On the other hand, in case I, because both the generation procedure and the deletion procedure play important roles on the network all the time, the deletion procedure can obviously reduce the increase of E by continuous and even more frequent (because values of t_g are larger than t_d) execution of the generation procedure, resulting in the sawtooth pattern. Therefore, we can predict that by strengthening effects of the relatively weaker procedure, or by weakening effects of the stronger procedure, E will appear with the sawtooth pattern more likely than smooth behavior. For example, for the graph represented by "o" in Fig. 8, it can be implemented by increasing g and d or decreasing t_g and t_d.

On the other hand, we find some similar effects of different parameters on E to case I: (i) E increases as g or d decreases, while E increases, as t_g or t_d increases, as shown in Fig. 8 and Fig. 9; (ii) E with $g=2$ is significantly larger than with $g=3$ and $g=4$ as in Fig. 8(a) and E with $d=0$ is also significantly larger than E with $d=1$ and $d=2$ in Fig. 8(b); (iii) E tends to become stabilized in the later part of the evolution process, as shown in Fig. 8 and Fig. 9. We find the reason for this is that generation and deletion of links rarely happen, so E remains unchanged after the network has evolved for a while. However, in case I, although the time average (over a period larger than t_d) of E becomes constant, this is because the two procedures happen continuously and balance each other after a while.

Having studied case II with degree-related e-mail contact, some similarities and some general laws of how different parameters affect the properties of the network as shown in case I with equivalent e-mail contact have been found.

V. CONCLUSION

In this paper, we have constructed and studied a novel model of evolving e-mail networks with nodes as users' address books and links as the records of e-mail addresses in the address books. Our model is close to the real world, while still keeping the simulations feasible. We apply the idea of the evolution to the network by generating and deleting links within specific time intervals. The model has been analyzed by using a probabilistic method and simulations. Two cases of e-mail contact sent from one user to other users at each time step have been considered in the simulations: (i) equivalent e-mail contact, (ii) degree-related e-mail contact. For both cases, we find that the statistical properties of this evolving e-mail network, such as average number of links, average path length, and clustering coefficient, are strongly affected by various parameter settings in simulations. We observe that the average number of links tends to increase or decrease depending on the values of generation and deletion threshold, and also the time interval to execute the generation and deletion procedures. Furthermore, in case I, the tendency of the average number of links and the sawtooth pattern reflects the relationship between generation and deletion procedures. Ultimately, the network reaches a stage at which the time average (over a period larger than t_d) of the number of links in the network becomes constant. In case II, the sawtooth pattern disappears and E becomes more smooth. Moreover, by analyzing simulation results in case I, we observe that for small-sized networks, the average path length between two nodes decreases as time elapses. With small values of average path length and high clustering coefficient, our evolving network exhibits small-world characteristic properties.

The ideas of evolving e-mail networks presented in our model can be also applied to model other real networks. This will be the subject of further work.

[1] P. Erdös and A. Rènyi, Publ. Math. (Debrecen) **6**, 290 (1959).
[2] B. Bollobàs, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, 2001).
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] A.-L. Barabàsi and R. Albert, Science **286**, 509 (1999).
[5] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[6] R. Albert, H. Jeong, and A.-L. Barab'asi, Nature (London) **401**, 130 (1999).
[7] B. Huberman and L. Adamic, Nature (London) **401**, 131 (1999).
[8] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[9] R. Albert and A.-L. Barabàsi, Rev. Mod. Phys. **74**, 47 (2002).
[10] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[11] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[12] A. L. Lloyd and R. M. May, Science **292**, 1316 (2001).
[13] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[14] M. E. J. Newman, S. Forrest, and J. Balthrop, Phys. Rev. E **66**, 035101 (2002).
[15] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).
[16] P. D. Wilde, BT Technol. J. **14**, 4 (1996).
[17] G. Caldarelli, F. Coccetti, and P. D. L. Rios, e-print cond-mat/0312236.
[18] A. Barrat, M. Barthèlèmy, R. Pastor-Satorras, and A. Vespignani, e-print cond-mat/0311416.
[19] S. Itzkovitz, N. K. R. Milo, G. Ziv, and U. Alon, Phys. Rev. E **68**, 026127 (2003).
[20] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. (McGraw-Hill International Book Company, New York, 1984).
[21] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[22] K. Klemn and V. M. Eguiluz, Phys. Rev. E **65**, 057102 (2002).